The Units of Speech Perception*

Ilse Lehiste

1. Introduction.

Speech perception is a vast topic that might be approached in several different ways. Much interesting work has been done recently with regard to models of speech perception. There is continuing interest in the question of categorical perception and the differences in perception depending on whether or not a listener is responding in the speech mode; related questions involve the role of lateralization in speech processing, and the relationship between speech perception and short-term memory. I have decided to limit the topic to a survey of recent work concerning the units of speech perception. It will occasionally be necessary to relate these units to units of production; likewise, it will be impossible to refrain completely from discussing certain speech perception models. However, I shall not attempt exhaustive coverage of these latter topics; in fact, it will not be possible to achieve exhaustive coverage even of the more limited subject. However, I hope to touch upon some of the more interesting theories and experimental findings at the several levels at which perception units may be established. I shall proceed from the smallest to the largest, starting with the perception of sub-phonemic phonetic differences and concluding with clause- and sentence-level units and their relationship to syntax.

2. The minimal units of speech perception.

2.1. Listening in the speech mode.

One of the problems in trying to establish what constitutes the minimal unit of speech perception is drawing a boundary between the perception of signals in a psycho-acoustic experiment (auditory processing) and the perception of signals in a speech mode (phonetic processing). It is well known that an identical physical stimulus may be perceived in two different ways, depending on the psychological setting. For example, the F_2 transitions of a synthetic CV syllable may sound as chirps of a bird or as glides in pitch, when presented out of context; provided with a following synthetic vowel, they signal the point of articulation of the consonant preceding the vowel (Liberman, 1970). The question is now whether listeners are capable of distinguishing subphonemic

phonetic detail while listening in a speech mode.

One of the characteristics of listening in a speech mode is the so-called categorical perception of phonemes. This means that a listener's ability to discriminate variations in the acoustic cue is much better at the boundary of phone classes than within the phone class (Liberman, Harris, Hoffman, and Griffith (1957); Liberman, Harris, Kinney, and Lane (1961); Stevens, Liberman, Öhman, and Studdert-Kennedy (1969)). Presented with a set of simulated CV- syllables in which F2 transitions are separated by the same frequency intervals, the listener groups the transitions according to the number of distinctive points of articulation employed in his language; within the range, adjacent sounds are classified as 'same', and crossing from one range to another, adjacent sounds are classified as 'different'.

There are some problems with categorical perception. In early experiments, it appeared to work well for consonants, but poorly for vowels. Categorical perception appeared to be associated with a discontinuity in articulation; in the case of vowels, there is no such articulatory discontinuity, which might explain a lack of categorical perception in vowels.

The problem has been recently re-considered by Chistovich and Kozhevnikov (1969-1970). It had been shown earlier (Fry, Abramson, Eimas and Liberman (1962); Stevens, Liberman, Öhman, and Studdert-Kennedy (1969)) that listeners are capable of distinguishing among a large number of stimuli (synthetic vowels) which are classified by them in the same phonemic category. This result could be interpreted in two ways. One interpretation is that phonetic images of vowels form a continuum; in hearing a vowel, the listener 'locates' the stimulus on the continuum by reference to certain articulatory target positions kept in memory. The other interpretation is that a listener is capable of remembering, for a certain time, not only the phoneme which has been selected on the basis of the heard stimulus, but also some spectral characteristics of the sound. If the two stimuli which are being compared prove to be different phonemes, subphonemic spectral information is discarded (Chistovich, Fant, de Serpa-Leitão, and Tjernlund (1966); Chistovich, Fant, and de Serpa-Leitão (1966); Fujisaki and Kawashima (1968)).

2.2. The subphonemic level.

The experiments discussed by Chistovich and Kozhevnikov showed that in certain cases, man is capable of perceiving subphonemic phonetic differences even while listening in a speech mode. This suggests that minimal units of perception may be found at a subphonemic level. A proposal to that extent has been recently made by Wickelgren (1969a, 1969b), who submits 'context-sensitive allophones' as candidates for the role of minimal perceptual units.

Wickelgren claims that sounds are determined by context in such a way that, for example, a /p/ preceded by /a/ and followed

by /i/ is uniquely determined as the kind of allophone that follows /a/ and precedes /i/, and such an allophone of /p/ is different from one that is both preceded and followed by /a/.

There are several problems connected with this model, some of which came up in connection with a recent study by Lehiste and Shockey (1971). In this paper, we explored the perceptual significance of transitional cues in one or the other of the vowels of a VCV sequence that are due to the influence of the transconsonantal vowel. Öhman (1966) had shown that the transitions from the first vowel in a VCV sequence to the intervocalic consonant depend on the quality of the second vowel. Likewise, there are differences in the transitions from the same consonant to the same second vowel that depend on the quality of the first vowel. In our study, we used taped VCV sequences (where V = /i æ a u/ and C = /p t k/) in which either the first or the second vowel was removed by cutting the tape during the voiceless plosive gap. Although the transitional cues were present, and were of the same kind and order of magnitude as those observed by Öhman, the listeners were unable to recover the missing vowels from these modified transitional cues.

According to Wickelgren's model, the context to which allophones are sensitive consists of one preceding and one following sound; thus a following /i/ in an /api/ sequence will not exert any influence on /a/, although it will influence the realization of /p/. The results of the experiment just reported might be considered supportive of Wickelgren's claim; although influence from the second vowel was physically present during the first vowel, that influence was perceptually insignificant. It would seem then that perceptually, the context to which allophones are sensitive is indeed limited to one preceding and one following sound.

There is another possible interpretation: the transitions both to and from the intervocalic consonant are part of the consonant; thus it cannot be claimed at all that V2 has affected V1, even though the transitions from V1 to C have been modified.

The first interpretation is supported by the vowel data, but contradicted by certain consonant data obtained in the same experiment (Lehiste and Shockey (1971)). Perceptually, the influence of the transconsonantal vowel was insufficient to recover the missing vowel; thus allophones seem not to be sensitive to non-contiguous context. However, the first vowel in a V1CV2 sequence is coded, according to Wickelgren's model, as $\#V_c$, the c being the same for different V1's regardless of the quality, or even the presence, of V2. In other words, to take a concrete example, the first /a/'s in /api/, /apa/ and /ap#/ should all be identically coded as $\#a_p$. It seems reasonable to assume that if the context-sensitive allophone is the minimal unit of perception, the context to which the allophone is sensitive should be perceptible. Thus the /p/ should be equally perceptible, i.e. equally recoverable, under all three conditions described above. Our experiments in consonant identification show extensive differences in identifiability between consonants that appear in final position as a

result of elimination of the second vowel on the one hand, or as a result of having been produced by the speaker as unreleased final consonants, on the other. Although the modifications of transitions to an intervocalic consonant due to the quality of a following vowel were not sufficient to recover that vowel, they did have an effect on the identification of the consonant when the second vowel was removed.

The stimuli used in the final consonant identification experiment should have been identical: the left-hand context of the intervocalic consonants and the unreleased final consonants was the same, and the right-hand context was effectively removed by elimination of the releases. If identification was based only on left-hand context, we would have obtained identical scores. Since the scores were considerably different, perception must have been influenced by the anticipatory effect of the right-hand context, manifested within the segment preceding the consonant.

As a digression, I would like to remark that the claim that sounds are not sensitive to noncontiguous context cannot be upheld anyway in the light of historical sound changes. There are numerous processes which affect sounds, e.g. vowels, across intervening consonants and vice versa. For example, in the so-called palatal umlaut that has occurred in Germanic languages, there must have been a stage at which the /a/ of, say, /api/ was clearly distinct from the /a/ of /apa/. Whether the intervocalic consonants were involved or not is a moot question; it is difficult to prove or disprove whether in the Germanic languages the intervocalic consonant was first palatalized and then lost its palatalization after transmitting it to the preceding vowel. There exist instances, however, in which a consonant that is otherwise susceptible to palatalization was not palatalized by a following high vowel under umlaut conditions.

Let us now return to the second possible interpretation: that the transitions are not part of the vowel at all, but part of the consonant. Then the vowel would consist only of the steady state. In principle, if a context-sensitive allophone is the basic unit of perception, the context to which it is sensitive should play a part in perception. In other words, if the transitions are part of the consonant, it should be possible to recover both the preceding and the following consonant in a C1VC2 sequence, given only the steady state of the vowel. We have not run such an experiment, but the recoverability of C1 and C2, in the correct order, from the steady state of the vowel seems implausible considering what is known of the effect of preceding and following consonants on vowel targets. For example, both a preceding and a following /r/ will lower the third formant of an interconsonantal vowel; but given only the steady state, it will not be possible to discover whether the lowering was due to left-hand or right-hand context.

Wickelgren's hypothesis thus seems to be in need of modification. It is clear that the effects of coarticulation reach beyond

contiguous sounds. On the other hand, the context is not always perceptually recoverable. It may be that the 'context-sensitive' allophones fit a production model better than a perception model. The physical modifications are undoubtedly there, but if the context of a context-sensitive allophone is not perceptible, it seems unjustified to assume that context-sensitive allophones are the basic units of perception.

Considering allophones as minimal units of speech perception is one way to approach a level of perception lower than the phoneme. Another is to consider phonemes as "bundles" of distinctive features, and to investigate perception at the feature level. There is no question but that certain features can be perceptually isolated from the "bundles" in which they appear; e.g., voicing can be extracted from the other characteristics of a voiced consonant. The fact that features can be responded to apart from the phonemes to which they belong supports the notion that the brain is capable of parallel processing of incoming information (Miller and Nicely (1955)).

Parallel processing has been discussed in detail in several recent publications (Chistovich and Kozhevnikov (1969-1970); Bondarko, Zagorujko, Kozhevnikov, Molchanov, and Chistovich (1968) (translated by I.L. (1970)); Liberman (1970)). In essence, it means that the same physical signal (e.g. a frequency change in the second formant) carries more than one kind of information (e.g. the phonetic value of a vowel and the point of articulation of an adjacent consonant). A corollary assumption is that it is difficult, if not impossible, to draw precise boundaries between acoustic segments in such a way that the first acoustic segment would contain no information regarding the perception of the second segment, and vice versa.

It will turn out that the first characteristic of parallel processing encourages us to seek the minimal units of speech perception at a level lower (in a certain sense) than traditional allophones, while the second characteristic leads to the conclusion that the smallest units of perception must be located at a higher level--the level of something like a syllable. Let us consider both propositions in somewhat greater detail, and relate them to the role of phoneme-sized units in speech perception.

But first of all I should remark that an assumption of parallel processing would partly save Wickelgren's 'context-sensitive allophones' as minimal units in speech perception: in effect, the perception process could operate with information contained in several time segments, and the problem of non-contiguous influence could be ignored. On the other hand, allophones would lose their unit-like character: their features, perceived separately and in parallel, would not necessarily be coterminous, and instead of phone-like units (which one assumes 'context-sensitive allophones' to be) we would be dealing with something like 'long components' (cf. Lehiste (1967-1970), discussing Harris (1944)).

The question of the perception of sub-phonemic phonetic detail leads back to the question of categorical perception. To the extent that listeners are capable of distinguishing between stimuli falling within the same phonemic category, we are dealing with the perception of sub-phonemic phonetic detail. Reference was made above to the work of Chistovich et al. (1966a, 1966b) which showed that listeners were able to make finer distinctions in vowels than those prescribed by their phonemic system. For evidence of sub-phonemic perception of a suprasegmental feature--duration--I should like to quote Lisker and Abramson (1971). In their experiments with the duration of voice onset time, one of the authors serving as listener distinguished five clear labeling categories, while the phonemic system of English would provide only two.

The differential perception of duration leads to the question of the perception of temporal segments in speech. Several phoneticians have expressed doubt concerning the possibility of perceptual segmentation of speech into units whose duration can be objectively established. It is, of course, known that acoustical signals are largely continuous; nevertheless, they also exhibit some drastic and abrupt changes. The continuous nature of the clues signalling the point of articulation has been used to argue that the minimal unit of perception is a unit of the order of a syllable (for a recent summary, cf. Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967)). On the other hand, continuous speech signals are perceived in ordinary listening as if they consisted of a sequence of discrete units (phonemes). The question is whether the boundaries of these units--or a modified version thereof--can in some way be associated with characteristics of the acoustic patterns. The basic question is thus whether it is possible to segment speech in a perceptually meaningful way.

The obvious place to begin is to consider signals that differ only in the duration of a segment, in such a manner that the differences in duration are not associated with any qualitative differences. The voice-onset-time experiments provide one such condition; they have shown both a possibility of categorical perception (which would serve as evidence for the phonemic level) as well as subphonemic perception (providing evidence for the ability of the ear to analyze duration in a phonetic rather than categorical manner). Further evidence is provided by languages with distinctive quantity.

It is a linguistic fact that in some languages the length of a vowel or consonant may have distinctive function. Experiments with synthetic speech (Lehiste (1970b)) show that listeners agree in a very high degree in assigning linguistic labels to stimuli that differ only in the duration of a vowel or consonant. This implies that listeners are able not only to compare the duration of two stimuli (such as the duration of a voiceless plosive gap), but also to match the stimuli with some kind of 'durational image', an abstract durational pattern characterizing a particular

word type. If a difference in duration of 10 milliseconds can switch 42% of the listeners from one category of linguistic response to another, the difference must be perceptually significant. Obviously it is impossible to tell, during the voiceless plosive gap itself, whether the plosive is qualitatively shorter or longer; the listeners must be comparing durations, which means that they must be using some fixed point of reference. I submit that at least in languages with distinctive quantity, abrupt changes in the manner of articulation serve as reference points with regard to timing judgments.

This is fully in accord with the notion that speech is processed in parallel; whatever the process by which the duration of one segment is compared with that of another (or with a stored 'durational image'), it can very well take place at the same time as the cues for point of articulation are processed which are extracted from the same acoustic signal (e.g. the same vocalic sound). In fact, all suprasegmental information must be processed in a similar way. For example, the presence of voicing serves to establish the voicedness of a vocalic sound at the same time as a possible fundamental frequency change taking place during the voiced segment may signal a distinctive lexical tone. I have discussed the perception of suprasegmentals in detail elsewhere (Lehiste (1967-70); Lehiste (1970a), and shall not elaborate any further on this topic within the present context.

There is additional, somewhat circumstantial, evidence of the importance of the manner of articulation in speech perception. In a study of the perceptual parameters of consonant sounds, Sharf (1971) established seven-point scales for duration, loudness, frequency, sharpness, and contact. Substantial numbers of significant differences were obtained only for duration comparisons based on manner of articulation (and for contact comparisons based on place of articulation; but since the contact parameter was specifically chosen to provide an indication of how well subjects related sounds to place of articulation, the latter finding appears unsurprising). In an earlier study, Denes (1963) showed that manner of articulation carries by far the greatest functional load in the English sound system, and suggested that the acoustic correlates of manner might be used for segmentation in automatic speech recognition systems.

Perception of duration thus appears associated with the perception of manner of articulation. Both represent perception of phonetic detail which may or may not be distinctive. The perception of such phonetic detail serves to substantiate the claim that the minimal elements of speech perception must be located at the subphonemic level, which may thus be considered as established.

2.3. The phonemic level.

The question is now whether the unit next in size is a phoneme-like unit or a syllable. The evidence for the psychological and

perceptual reality of phoneme-like units has been summarized by Chistovich and Kozhevnikov (1970). Savin and Bever (1970) have argued for the "non-perceptual reality" of the phoneme. Let us review the arguments of Chistovich and Kozhevnikov first.

Much of the evidence for phoneme-like perceptual units comes from studies of categorical perception (cf. above). To the extent that the categorical perception idea is valid, the psychological reality of phonemes as perceptual units must be accepted. There is a connection between categorical perception and the motor theory of speech perception; both seem to apply better to consonants than to vowels (or to other signals of a continuous nature) (Liberman (1957); Stevens (1960); Liberman, Cooper, Harris, and MacNeilage (1962); Lane (1965); Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967); Studdert-Kennedy, Liberman, Harris, and Cooper (1970)). Chistovich and Kozhevnikov (1969-70) have shown, first, that vowels are also perceptible in a categorical fashion. Since the articulatory process involved is continuous rather than discontinuous, this would argue against the motor theory. Second, they suggested that the number of categories in vowel perception may be larger than the number of traditional phonemes in the language; and further, that a listener is capable of remembering for a certain time not only a phoneme, but what they call 'timbre description'--subphonemic phonetic detail, which makes it possible to make distinctions within a category. The authors call their perceptual categories 'psychological phonemes'. It has been shown, for example, that Russian subjects classify [ɨ] and [i] as different psychological phonemes, although they are never encountered in the same environment and thus may be considered as constituting allophones of a single phoneme. Vowels between hard and soft consonants were classified by Russian subjects as belonging to different sound types, although they would again constitute positionally conditioned allophones according to classical phonemic theory.

Savin and Bever (1970) studied the order in which listeners make decisions at the phonemic and syllabic levels in the course of speech perception. Their method was to ask a listener to monitor a sequence of nonsense syllables for the presence of a certain linguistic unit, either a phoneme or a syllable, and to respond (by releasing a telegraph key) as quickly as possible when he had heard it. The target was a complete syllable (e.g. "bæb", "sæb") or a phoneme from that syllable: the syllable-initial consonant phoneme for some subjects (e.g. /b/ or /s/) and the medial vowel phoneme for other subjects (e.g. /æ/). Subjects responded more slowly to phoneme targets than to syllable targets (by 40 msec for /s-/, 70 msec for /b-/ and 250 msec for medial /æ/). Savin and Bever interpret these results as supportive of the view that phonemes are identified only after some larger linguistic sequence (e.g. syllables or words) of which they are parts. The reality of the phoneme, the authors say, is demonstrated independently of speech perception

and production by the natural presence of alphabets, rhymes, spoonerisms, and interphonemic contextual constraints.

These results do not disprove the existence of a phonemic level of perception, and therefore the title of the paper by Savin and Bever ("The nonperceptual reality of the phoneme") appears somewhat misleading. Before the general conclusion is accepted, one would like to see what the reaction times to final consonants are, i.e. whether subjects would respond more slowly to a final /-b/ than to the syllable /sæb/. While not directly comparable to the reaction time experiments carried through by Fry (1970, to be discussed below), the results of Savin and Bever are sufficiently different from those of Fry to suggest additional studies.

It seems that a level of perception at which phoneme-like units are responded to should be recognized; it remains to relate it to the other levels of perception for which evidence has likewise been provided by studies of speech perception.

3. Higher-level units of perception.

3.1. Unitary perception of sequences of segments.

The parallel processing of speech signals is compatible with the suggestion that the minimal unit of perception must be of the order of a syllable (Savin and Bever (1970); Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967)). There is a good bit of evidence that the ear is particularly well suited to the perception of changes in acoustic parameters rather than their steady states (Abbs and Sussman (1971)). Without going into details, let me just recall the experience of most researchers who have synthesized isolated vowels: produced on a monotone, the vowels frequently seem to occupy a borderline between speech-like and nonspeech-like stimuli, while the imposition of a fundamental frequency glide shifts the listener clearly into the speech mode. It is also well known that the majority of point of articulation cues of consonants are manifested in adjacent vowels. It seems thus reasonable to look for higher-level units of perception beginning with sequences of two speech sounds. The first major problem involves the perception of sequential order.

Wickelgren's idea of context-sensitive coding could certainly explain the correct perception of sequential order; but the notion of parallel processing, which seemed essential for upholding that theory, appears to be incompatible with the decoding of order from simultaneously received feature cues. The perception of temporal order is a vast topic, deserving a review on its own; I shall restrict myself in this survey to a few recent experiments which shed some new light on the problem.

The mechanisms employed in the perception of consonant clusters have been investigated in a series of experiments by Bond (1971) and Day (1970a, 1970b).

The study by Bond (1971) deals explicitly with the perceptually unitary nature of consonant clusters. Bond studied 15 pairs of English words which differed from each other only in the order of obstruents in the cluster. The pairs /ps-sp/, /ts-st/ and /ks-sk/ were all represented five times (some examples: task-tax, lisp-lips, coast-coats). The words were produced by a male native speaker of English; randomized listening tests were constructed, in which the signal was degraded by addition of white noise. 19 subjects took the listening test, writing down what they heard. Five of the subjects took the test a second time, producing a spoken response (a repetition) to each stimulus. These subjects' responses were analyzed for reaction time in addition to being scored for correctness. It was found that reaction time was consistently faster for correct than for incorrect responses; but the pattern of confusions for written responses and spoken responses was essentially the same. It was further found that reversal errors were the most common errors. Bond argues from this that minimal perceptual units must be larger than the phoneme. If consonant clusters were perceived phoneme by phoneme, there is no reason for the listener to reverse the order. To be sure, the listener may occasionally be forgetful; but there is no reason to suppose that he would be more likely to forget the order of the consonants than to forget one of the consonants. Since reversal errors were much more common than substitution errors, some special perceptual mechanisms must be postulated for the perception of consonant clusters. Bond's findings thus confirm a suggestion made by Neisser (1967), according to which a listener gradually learns to distinguish a cluster like /ts/ from a cluster like /st/, rather than perceiving a sequence of /t/ followed by /s/, or /s/ followed by /t/. Clusters of this type thus seem to constitute a perceptual unit.

Day (1970a) studied phonemic fusion in dichotic listening, in which listeners received two speech stimuli at the same time with various relative onset times. The stimuli differed in their initial consonants (e.g. /bæŋkət/ and /læŋkət/). On some trials, either /bæŋkət/ or /læŋkət/ led by 25, 50, 75, or 100 msec; on other trials, both stimuli began at the same time. Subjects reported hearing /blæŋkət/ regardless of which consonant led. When specifically asked to judge the temporal order of the initial phonemes, most subjects reported hearing /b/ first, no matter whether /b/ or /l/ actually led. Day concludes that instead of processing temporal order in an accurate fashion, subjects responded to the stimuli according to the constraints imposed by the phonological system of English. In English, stop + liquid clusters are permissible in initial position, but liquid + stop clusters do not occur. The responses thus clearly imply the presence of a linguistic level of processing.

A similar study was carried out with reversible clusters (Day (1970b)). Since there are no reversible clusters in English in initial position, a final cluster was selected. The stimuli

were /tæs/ and /tæk/, whose fusion would yield acceptable English words in either order, viz. /tæsk/ and /tæks/. All trials were dichotic pairs, consisting of /tæs/ to one ear and /tæk/ to the other ear. The onsets of the syllables were aligned over a wide range of values: stimuli either started at the same time, or one or the other stimulus led in steps of 5 msec to a 100 msec lead.

In contrast with the nonreversible case, temporal order judgment was very good when the cluster could occur in either order in the language. One of the temporal orders (/ks/) was somewhat more preferred. Day suggests that this may be due to the fact that the acoustic shapes of stop consonants undergo greater changes as a function of context than do fricatives; thus the acoustic shape of /k/ in /tæk/ may be more important than that of the /s/, to the extent of biasing the perceived order of the two phonemes. (I would suggest that segmental duration may have played a perhaps decisive part. The stimuli were synthesized with equal duration given to /æ/ in both /tæk/ and /tæs/. In actual speech, /æ/ would be longer before a fricative; thus listeners may have been biased toward a /tæks/ response by the relative shortness of the /æ/).

In a further experiment, subjects were asked to decide which ear led, rather than which phoneme. Performance on the ear task was much better: subjects were highly accurate, even though they were language-bound on the phoneme task.

The difference between the results obtained with nonreversible and reversible clusters is explained by Day as follows. Two general levels of processing are postulated: a linguistic level and a nonlinguistic level. Both operate in normal listening situations, but the linguistic level appears to be prepotent: it can effect selective loss of information obtained from the nonlinguistic level. Correct temporal order may be represented in the system at some point in time, but later stages of processing mold this information to conform to the linguistic structure of the language. Hence nonlinguistic information, concerning acoustic shape and temporal order, may be lost or ignored. Day suggests that temporal order information is lost only after it enters higher stages of linguistic processing.

3.2. Primary processing and linguistics processing.

Day called the two levels of speech processing which her experiments had isolated linguistic and non-linguistic. It appears, however, that both levels have to be further subdivided. Even at the non-linguistic level, there is a difference in perception depending on whether one is listening in the "speech mode". Evidence for this is available from many sources, among which are laterality studies (Studdert-Kennedy and Shankweiler (1970); Day and Cutting (1970)). I would like to call the processing of an auditory signal in the speech mode "phonetic processing". Attempts to separate auditory and phonetic modes of

processing have been recently discussed by Fujisaki and Kawashima (1969) and by Pisoni (1971). The linguistic level suggested by Day could perhaps be called the phonological level of speech processing. At this level, information available to the listener about the phonological structure of the language (e.g. information concerning permissible sequences) is interposed between primary recognition and perceptual decision. The experiments of Chistovich et al. (1966a, 1966b) regarding the mimicking and perception of vowels show the possibility of separating the phonetic and phonological levels of perception, as do the experiments in the perception of reversible and non-reversible clusters by Day.

There are higher levels within the linguistic level of processing, and some attempts have been made recently to explore them experimentally. A very intriguing set of experiments by Fry (1970) deals with reaction time to monomorphemic and bimorphemic words that are identical as to their phonemic composition. Fry used the minimal pair lacks/lax, serving both as speaker and listener. Responding 100 times to the randomized stimuli, he made only 2 wrong responses to 50 occurrences of lax, and likewise only two errors in responding to lacks--a result surprising to Fry, who had not expected a subject to be able to respond consistently to the difference between the two items. The mean reaction times were 557 msec for lax and 518 msec for lacks, a difference that just misses significance at the .05 level of probability. Fry considers it worth noting that the direction of the difference points to a longer reaction time to the monomorphemic word.

Fry also tested the reaction time to longer sequences differing in the presence and absence of a word boundary. The items were the two sentences It's a sign of temporizing and It's a sign of temper rising, which are segmentally identical in Fry's pronunciation. There were six errors in the perception of 50 presentations of temporizing and 3 in the case of 50 presentations of temper rising. Mean reaction times (measured from the beginning of the syllable /tem/ in each case) were 711 msec for temporizing and 858 msec for temper rising, a difference which was significant below the .01 level of probability. The item containing the word boundary thus took significantly longer to produce a response, although the difference in duration between the two items was negligible (30 msec in a total of 1430 msec).

Fry's starting assumption had been that processing time increases with the complexity of the task. The results of the experiment with sentences support this view; the two sentences differ in their syntactic structure, and it is quite probable that the syntactic level of processing was involved in addition to primary processing. However, the results of the lax - lacks experiment seem to imply that a monomorphemic word presents a more complex task than a bimorphemic one. This appears counter-intuitive; and there might be alternative explanations to Fry's

findings. If the results should be substantiated by further experiments, it might be assumed that a bimorphemic word contains more information than a monomorphemic word and therefore can be processed faster. If additional data should show that the effect observed by Fry may have been due to chance, it might be concluded that there exists no separate morphemic level of linguistic processing.

Such experiments were in fact carried through by Bond (1971). Bond used ten minimal pairs, each pair consisting of one monomorphemic and one bimorphemic word of the same phonemic shape. Each pair of words composed a sub-list, within which the two words were recorded in random order, each word being produced ten times. Care was taken to insure that the speaker intended the 'right' word every time. 29 listeners took the test, which consisted of 200 stimuli. Reaction times and correct scores were obtained by techniques similar to those used by Fry.

The overall scores indicated that subjects were not able to identify the words correctly at levels significantly above chance. The mean scores ranged from 45.1% for lax - lacks to 55.4% for lapse - laps. When the responses of the subjects to each production were analyzed, however, it was found that subjects were very consistent in their responses to some of the test items. Significant scores (at the .02 level) were obtained for three items in the 20 productions of members of the pair bard - barred (15.4%, 84.6% and 15.4% correct), and one item each in the pairs wade - weighed (100% correct), lax - lacks (18.2% correct), baste - based (85.7% correct) and mist - missed (100% correct). As the scores show, while the subjects could be highly consistent in agreeing on a particular response, they did not necessarily identify the word correctly; the identification scores for utterances on which the subjects agreed on one response were still at chance level (57% correct).

There was no significant systematic difference in reaction time between correct and incorrect responses. There was, however, some tendency for reaction time to be shorter to the bimorphemic word, as Fry had discovered; the differences were not statistically significant.

This cannot be considered supportive of Fry's findings, because reaction time differences become meaningful only if the subjects can identify the words correctly, which was not the case with Bond's subjects. Bond explains the high degree of agreement shown by the subjects in response to some of the stimuli as follows. Faced with the task of the experiment, listeners develop a strategy for making use of fine phonetic detail (duration, spectral characteristics of /s/ etc.). In this manner they arrive at some consistent labelings. But since the identifications based on this strategy are equally likely to be correct or incorrect, the strategy cannot be considered to be part of ordinary speech perception.

Within the framework developed in this paper, I would propose that we are dealing with phonetic processing rather than linguistic

processing. The perception of fine phonetic detail is certainly documented by Bond's results, but this information plays no part in establishing a possible morphological level within linguistic processing.

While the morpheme level evidently has to be rejected as a level of processing within the level of linguistic processing, it might be inquired whether a word constitutes a perceptual unit at some level. Fry's reaction time experiments provide some evidence that the word is certainly not the minimum unit of perception. In testing reaction times to 18 contrasts like bid-big, or begin-began, Fry found that in only three cases did the mean reaction time exceed the total duration of the stimulus. In most cases, subjects had no difficulty whatever in responding before a word or syllable were complete. The processing mechanism was evidently capable of dealing with segments smaller than the whole syllable or word.

Whether the word constitutes a perceptual unit does not emerge from Fry's experiment with sentences containing the items temporizing - temper rising, since in examples of this kind it is impossible to separate lexical differences from syntactic ones. However, certain techniques have been developed within the past ten years for studying the perception of syntactic units, and the rest of the paper will deal with perception at this level.

3.3. Perception of syntactic units.

To a large extent, recent studies of sentence-level perceptual units go back to a seminal paper by Ladefoged and Broadbent (1970). In the research on which the paper is based, Ladefoged and Broadbent presented a series of tape-recorded sentences to various groups of listeners. During each sentence, a short extraneous sound (a "click") was present on the recording, and listeners had to indicate the exact point in the sentence at which the click occurred. Errors were large compared to the duration of a single speech sound; Ladefoged and Broadbent concluded that the basic unit of perception is larger than a phoneme, and that the listener does not deal with each sound separately but rather with a group of sounds. Subjective location of clicks, as reported by the subjects, differed from their objective location according to a regular pattern; Ladefoged and Broadbent argue that the points toward which the clicks were displaced constituted boundaries of perceptual units.

Fodor and Bever (1965) used the same technique to investigate the hypothesis that the primary units of speech perception correspond to the constituents of which a sentence is composed, i.e. the more abstract segments revealed by a constituent analysis of the sentence provided by the grammar of the language. Fodor and Bever found that clicks were attracted toward the nearest major syntactic boundaries in sentential material. The number of correct responses was significantly higher in the case of clicks located objectively at major boundaries than in the case of

clicks located within constituents. Fodor and Bever consider these results supportive of the view that the segments marked by formal constituent structure analysis do in fact function as perceptual units, and that the click displacement is an effect which insures the integrity of these units: the units resist click intrusion.

In a subsequent study, Garrett, Bever and Fodor (1965) attempted to determine whether the earlier results should be interpreted as reflections of the assignment of constituent structure during the processing of sentences, or were rather effects of correlated acoustic variables (such as pause and intonation) which tend to mark constituent boundaries in spoken language. They constructed and recorded pairs of sentences for which some string of lexical items was common to each member of a pair. The common portions of each pair were made acoustically identical by cross-splicing, i.e. by splicing a recorded version of a portion of one member of the pair to the opposite member of the pair. When a spliced version is paired with a copy of the original recording,

(Example: A. (In her hope of marrying) (Anna was surely impractical)
B. (Your hope of marrying Anna) (was surely impractical).)

there are two sentences in which part of the acoustic material is identical, but for which the constituent boundaries are different. The results showed that exactly the same acoustic signal was responded to differently in every case, and the differences were uniformly as predicted by the intended variation in the constituent structure.

Bever, Lackner and Stolz (1969) further tested the hypothesis that the perceptual segmentation of speech depends on transitional probabilities. The fact that clicks are subjectively located at boundaries between clauses might be a reflection of the low transitional probability between clauses rather than a demonstration that syntactic structure is actively used to organize speech processing. In this experiment, subjects were asked to indicate the subjective location of clicks placed in sentences which differed in terms of transitional probabilities between clauses. It was found that high-probability sequences within clauses attract clicks, while low-probability sequences do not. The authors interpret these results as indicative that transitional probability has different effects within and between clauses and thus is not a general mechanism for the active segmentation of speech.

In another set of experiments, Bever, Lackner and Kirk (1969) found that within-clause phrase structure boundaries do not significantly affect the segmentation of spoken sentences, and that divisions between underlying structure sentences determine segmentation even in the absence of corresponding

clause division in the surface phrase structure.

In most of these studies, subjects were ostensibly involved in only one task, namely click localization; but in fact they were performing a far more complex assignment. They had to listen to a sentence, pay attention to the click, remember the sentence, write it down, remember the click location, and mark that on the written version of the sentence. The sentences were usually quite long; it seems obvious that we are dealing here with a complex interaction of perception and memory. Techniques used up to this point did not attempt to separate the effects of memory and perception.

Abrams and Bever (1969) attempted to minimize the effects of memory by giving the subjects a different task: pressing a key in response to a click. In a second presentation of the test sentences, subjects had to write the sentences and locate the click as before. Reaction times were thus obtained in addition to click localization data.

The results turned out somewhat ambiguous. Abrams and Bever had expected that clicks objectively occurring in clause breaks should receive faster reaction times than clicks in any other location. This turned out not to be so. There was also no systematic interaction between reaction time and subjective click location. Reaction time to clicks before clause breaks was affected by clause length and by familiarity with the sentence more than the reaction time to clicks after clause breaks. According to Abrams and Bever, this indicates that syntactic structure does systematically modify attention during speech perception. In sentences, the clause is a natural unit for internal perceptual analysis. During clauses one listens to the speech and nonspeech stimuli; at the end of clauses one encodes perceptually what was just heard. Accordingly, a click at the end of a clause is responded to relatively slowly, since it coincides with the point of internal perceptual analysis of the preceding sentence. At the beginning of a clause, a click is reacted to quickly because it conflicts with relatively little internal perceptual processing.

Abrams and Bever suggest further that the attentional system tapped by the reaction-time measure is distinct from the behavioral process which produces the systematic errors in click location. Immediate reaction time interacts with the process of developing the internal perceptual organization of speech. Listeners first organize the speech into major segments, then they relate the speech and click temporally. It is this latter process that maintains the integrity of the speech units as revealed in the location of clicks.

In another study, Bever, Kirk and Lackner (1969) tried to avoid conscious participation of the listeners altogether by measuring their galvanic skin response to shocks. In this experiment, subjects heard sentences in one ear, during which a brief shock was administered before, in or after the division between two clauses. The galvanic skin response to shocks

objectively at the end of a clause was larger than the response to shocks at the beginning of a clause. Bever, Kirk and Lackner view this as confirmation of the hypothesis that the syntactic structure of a sentence can influence systematically the change in skin resistance in response to a mild shock presented during the sentence.

An independent effect was that galvanic skin response to shocks at the end of a clause decreased as a function of clause length; responses to shock at the beginning of a clause were relatively unaffected by the length of the preceding clause. According to the authors, this supports the claim that listeners respond to the syntactic structure of speech as they hear it.

Fodor and Garrett (1971) revised the earlier view that click location is affected only by major constituent boundaries. Under appropriate conditions (when a listener is given more than the usual amount of time to consider a sentence), minor boundaries were found to affect click location. Fodor and Garrett suggest that assignment of minor constituent boundaries is a relatively late operation in the processing of sentences. If the listener has a chance for developing a more fine-grained analysis of the sentence containing a click, effects of minor constituent boundaries on click location are increased.

The series of studies just reviewed thus presents the following claims: listeners use grammar actively to impose syntactic structure on the speech stimulus as they hear it. Listeners respond in terms of the underlying structure of the sentence rather than its surface structure. Acoustic cues alone do not determine the boundaries of perceptual units.

Certain of these findings have been challenged in several recent studies. Abrams and Bever (1969) had found that subjects did not react faster to clicks placed in major constituent breaks than to clicks within the constituents. Holmes and Forster (1970) found exactly the opposite: reaction times to clicks at the major syntactic break of the sentence were faster than reaction times to clicks not at a break. This confirmed their hypothesis that processing load is a function of the surface structure of sentences, and that it decreases at major constituent boundaries.

The second result of the study by Holmes and Forster is likewise in direct contrast to the findings reported by Adams and Bever: reaction times were slower when the click was in the first rather than in the second half of the utterance. Holmes and Forster interpret this result likewise in terms of differential processing loads. It is obvious that these results place in question the conclusions drawn by Abrams and Bever from their data.

Chapin, Smith and Abrahamson (1972, in press) produce a detailed critique of Bever, Lackner and Kirk (1969) who had claimed that underlying structure sentences are the primary units of immediate speech processing. Chapin, Smith and Abrahamson found that clicks were attracted to major surface constituent boundaries, even when these did not coincide with the boundaries

of underlying structure clauses. Another finding was that clicks are attracted to preceding constituent boundaries. This suggests an overriding perceptual strategy in speech processing: the listeners attempt to close constituents of the highest possible level at the earliest possible point.

Bond (1971) studied both click localization and reaction time, testing the hypothesis that subjects segment an incoming sentence on the basis of stress and intonation patterns. Reaction time is then predicted to be shorter to clicks between phonological phrases, and longer to clicks within phonological phrases; it is also expected to be different to clicks located in stressed syllables, as compared to clicks placed in unstressed syllables.

When reaction time to clicks in stressed and unstressed syllables was compared, it was found that reaction time was significantly faster to the click located in an unstressed element, either in the consonant preceding the unstressed vowel or in the unstressed vowel itself. Subjects were much more accurate in locating a click when it occurred in a stressed vowel than when it occurred in a consonant or in an unstressed vowel (correct scores 46% vs. 12%). Clicks were thus much less likely to be 'attracted away' from stressed vowels than from unstressed vowels; the error responses, however, were in the direction toward major boundaries.

Reaction time was also examined on the basis of an 'intonation phrase', i.e. any phrase that was demarcated by a clear intonation curve. Reaction time was found to be progressively slower as the click occurred further into the intonation phrase; thus there is a correlation between reaction time and the position of the click within an intonation phrase.

Bond suggests that in sentence perception, the listeners segment the sentence into phrases defined on the basis of stress and intonation; they then process the sentence further, to arrive at a syntactic analysis. Reaction time is apparently sensitive to initial segmentation, while click localization is sensitive to the final analysis.

3.4. The Role of Stress in the Perception of Sentence-Level Units.

Bond's study did not attempt to separate the parts played by stress and intonation. I conducted an experiment, described below, to investigate further the role of stress in click localization.

The purpose of this experiment was to explore the role played by suprasegmental features, especially stress, in the analysis of an incoming sentence. If the assumption is true that linguistic processing presupposes phonetic processing, it stands to reason that stress and intonation are not ignored by a listener in the perception of a sentence. This, as may be recalled, has been more or less generally assumed since the 1965 paper by Garrett, Bever and Fodor (cf. above).

It was decided to place clicks in identical positions within a sentence, varying the stress in such a manner that the words

within which clicks occurred would appear both with and without stress, all other factors being equal. If listeners react differently to clicks placed in the same position under different stress conditions, the role of suprasegmental factors in perceptual processing will be confirmed.

In order to control stress and click placement precisely, the experiment was carried through with synthetic speech. The stimuli were produced at the Bell Telephone Laboratories using the following technique. A normal utterance was analyzed by a formant-tracking program (Olive (1971)). The automatically tracked formants and fundamental frequency were later modified by hand; changes in time, formant structure, and fundamental frequency were produced by a suitable computer program. The program allows the researcher to specify the frequencies of the three formants, the fundamental frequency, and the overall amplitude at each 10 msec sampling period. Specific changes that were made will be described below. The re-synthesis was produced by a digital hardware synthesizer (Rabiner et al. (1971)). The entire process was controlled by a Honeywell DDP 224 computer (Denes (1970)).

The experimental technique used in the experiment differs from earlier methods in several ways. In most previous experiments, clicks had been recorded on the second channel of a two-track tape recorder, and the stimuli had been presented to listeners dichotically through headphones. Dichotic presentation introduced into the experimental situation a whole array of complicating factors, including competition between speech and nonspeech in relation to hemispheric specialization (Day and Cutting (1970)), and the problem of right- or left-handedness of the subjects. To avoid these probably unnecessary complications, the stimuli were recorded on full-track tape, with clicks introduced synthetically within the recording, and were presented to listeners over a good-quality loudspeaker in a sound-treated environment. It should be recalled that Ladefoged and Broadbent had likewise used a loudspeaker in their original experiment reported in 1960.

In most earlier experiments, listeners were required to write down the sentence that had been presented, and to indicate the position of the click on their own transcription. As was mentioned above, this technique introduces a memory component into the picture whose magnitude is difficult to estimate. It has been known for some time that the human short-term memory has a capacity of something like seven syllables (Miller (1956)). Memory units have been studied intensively by Johnson (1970), who found the 'chunks' of recall to be approximately the same size. In a recent paper, Gamlin (1971) has shown that subjects matched for intelligence may differ in their short-term memory capacity, and that low short-term memory subjects structure sentences differently than high short-term memory subjects. Gamlin suggests that low short-term memory subjects may be forced by their memory limitations to process sentences into smaller syntactic units.

Clearly most of the test sentences used in earlier click experiments have been long enough to overtax the short-term memory; thus it is entirely possible that the results confuse the sentence processing strategies with memory strategies.

The way chosen to eliminate the memorization problem was to use only one sentence with which the listeners became familiar during the introduction to the test, and to provide the subjects with written versions of the sentence. This represents again a return to the Ladefoged-Broadbent (1960) technique. In that study, subjects were presented both with unknown sentences over headphones, and with sentences that were written out and read out before the stimuli that contained the clicks were played over a loudspeaker. Ladefoged and Broadbent found that prior knowledge of the content of the sentence did not affect accuracy.

The sentence chosen for the experiment was one used by Bever, Lackner and Kirk (1969) in the experiment which provided the basis for their claim that the underlying structures of sentences are the primary units of immediate speech processing. The sentence, together with the phrase structure assumed by Bever et al., is as follows:

If (you (did ((call up) Bill))) (I (thank you (for (your trouble))))

Bever et al. placed clicks in the major clause break, in the middle of each of the two words immediately preceding the major break, and in the middle of each of the two words immediately following the major break. Separate results are not reported for this sentence, but one may assume the general conclusions to be applicable, i.e. that the boundary after Bill attracted clicks, while boundaries within the two clauses had no consistent effect on the subjective location of clicks.

The sentence was synthesized by the procedure described above. The sentence was produced by a male speaker with no special emphasis on any word and without any pauses. After resynthesis, the pitch of the sentence was changed to monotone at 100 Hz. Stress was then simulated on each of the four words did, Bill, I, and thank. This was done by time expansion and by introducing a pitch inflection on the appropriate word. The values of the parameters are specified by the program at 10 msec intervals. In time expansion, the number of sampling intervals is specified to which a given word is to be expanded, and the program interpolates the values of the parameters proportionately. The expansion factors had been obtained previously by comparing the durations of stressed and unstressed versions of the test words in different productions of the sentence; they were 25/33 for did, 32/58 for Bill, 16/34 for I and 31/42 for thank.

The fundamental frequency contour applied to the test word started at 100 Hz, rose to a peak of 111 Hz, and dropped back to 100 Hz. The peak of the contour was placed at the point of occurrence of the fundamental frequency peak in a normal stressed production.

Clicks were produced by setting formant frequencies to 1 for one sampling period and introducing random noise through the formants at an intensity equal to that of the strongest vowel. The duration of the clicks was 10 msec. Clicks were placed before, within and after each of the four words; the clicks within words were located at the pitch peak. With the method of time expansion used in the study, the clicks remained in precisely the same position relative to the word under both stress conditions. A table of click placements and stress conditions is given below.

TABLE 1
SURVEY OF CLICK PLACEMENT

Stressed word	Test word	Click placement relative to test word		
		Before	Within	After
Did	did	x	x	x
	Bill	x	x	x
	I		x	x
	thank		x	x
Bill	did	x	x	x
	Bill	x	x	x
	I		x	x
	thank		x	x
I	did	x	x	x
	Bill	x	x	x
	I	x	x	x
	thank		x	x
thank	did	x	x	x
	Bill	x	x	x
	I		x	x
	thank		x	x

Two comments should be added. In order to simulate stress on I, a glottal stop (with a duration of 17 sampling periods, i.e. 170 msec) was inserted before I. In the sentence in which I carried simulated stress, two click placements were used for the sequence Bill, I: a click was placed in the last frame of Bill, immediately preceding the glottal stop, and in the first frame of I, immediately following the glottal stop. In other instances, only one click placement was used between words. This is true also of sequences of Bill, I (i.e. the major clause break) in all other cases in which I was not stressed, including those in which Bill carried simulated stress.

The first part of the listening test was designed to check the effectiveness of the stress simulation. A set of ten randomized sentences was prepared, containing two productions

each of the test sentence produced on a monotone (and without time expansion, i.e., without stress simulation), and two sentences each with stress placed respectively on did, Bill, I, and thank. (The sentences contained no clicks.) The listeners were asked to underline the stressed word. The results are presented in the following table.

TABLE 2
SUBJECTIVE PLACEMENT OF STRESS, DEPENDING ON STRESS SIMULATION
Scores in per cent

	If	you	did	call	up	Bill,	I	thank	you	for	your	trouble.
Monotone	4	4	12	16	16	28		10				10
Stress on did		4	90	2				4				
Stress on Bill			6	2	2	84		6				
Stress on I			2	2		4	92					
Stress on thank	2		2			2		94				

As may be seen from the table, the syllables on which simulated stress was placed were overwhelmingly accepted as being stressed. The neutral sentence provided two surprises. I had expected the word did to be judged as stressed, since it is lexically marked as emphatic; however, there was a wide scatter of responses, and the word judged relatively most frequently as stressed was the word Bill. It will be reported later that this word behaved in an unexpected way in other respects too. Whether its position before the clause break is in any way connected with this behavior has to remain a matter of conjecture; further experimentation is clearly needed to solve the problem.

After the first part of the test, some examples of sentences containing clicks were played to the listeners, and instructions were given to draw a slash line through that part of the sentence that contained the click. Subjects were informed that clicks may occur between words or within a word. Sample sentences with slashes were provided on the handout. The subjects then proceeded to the main part of the test, which contained the 41 stimuli in two different randomizations (for a total of 82 stimuli), balanced in such a way that each stimulus occurred once during the first half and once during the second half of the test. The whole test took approximately twenty minutes to complete. The test was

administered singly or in small groups to 25 listeners, mainly graduate students and staff members of the Department of Linguistics of the Ohio State University. The results consist of 50 judgments per stimulus, for a total of 4100 judgments. The results of the listening tests will be presented with reference to Tables 3, 4, and 5. The question of correct identification will be discussed first.

The evidence for the listener's analysis of the sentence in terms of underlying structure units had been largely derived from subjective localization of clicks at major syntactic boundaries. Specifically, it had been claimed that clicks objectively at such boundaries were correctly located more frequently than clicks placed elsewhere, and that clicks placed elsewhere had a strong tendency to migrate toward the major syntactic boundaries. This experiment contained sentences in which clicks were placed at various boundaries, including the major clause boundary. The per cent correct identification of click location at various boundaries was as follows:

If (you (did ((call up) Bill))) (I (thank
24.0 40.0 51.5 27.6 41.5 16.5

you (for (your trouble))))

The total number of clicks correctly identified between Bill and I was 69 out of a possible 250 (5 sentences), or 27.6%. The total number of clicks objectively placed in the boundary, but subjectively shifted elsewhere, was 181, or 72.4%. Most of these clicks were attracted into the following word, i.e. into I. When I was unstressed, it attracted 37 clicks away from the boundary (from 150 possibilities, 3 sentences), and when it was stressed, 71 (from 100 possibilities, 2 sentences). As far as attracting clicks objectively located elsewhere, there were 150 such cases out of a possible 1800 (36 sentences), which amounts to 8.3%.

It must be concluded that the results of this experiment do not support the claim that the major syntactic boundary attracts clicks.

Table 3 presents the average correct scores for the subjective location of clicks objectively placed in stressed and unstressed productions of the words did, Bill, I, and thank. The unstressed scores combine stresses on the three other words; e.g. unstressed did combines scores for instances in which stress was simulated on Bill, I, and thank. A study of the scores reveals a number of regularities. There is a common pattern for the words did, I, and thank, while Bill shows a highly divergent pattern. Table 4 gives the average scores of the three words with similar behavior.

TABLE 3
CORRECT SCORES (PER CENT)

Word	Objective click placement		
	Before	Within	After
did, stressed	16.0	68.0	62.0
did, unstressed	26.7	56.0	32.7
Significance of difference*	> .10	> .10	< .01
Bill, stressed	56.0	38.0	24.0
Bill, unstressed	50.0	64.7	32.7
Significance of difference	> .10	< .01	> .10
I, stressed	16.0	40.0	66.0
I, unstressed	34.7	24.0	33.3
Significance of difference	< .05	< .10	< .001
thank, stressed	28.0	78.0	22.0
thank, unstressed	46.0	43.3	14.7
Significance of difference	< .05	< .001	> .10

*See Spiegel (1961, p. 171).

TABLE 4
CORRECT SCORES FOR DID, I, AND THANK (IN PER CENT)

Word	Objective click placement		
	Before	Within	After
Stressed	20	62	50
Unstressed	35.8	41.0	26.9
Significance of difference	< .10	< .05	< .001

In unstressed versions of did, I, and thank, clicks placed before the word tended to be identified more correctly than clicks placed in analogous position in stressed words. The difference is significant at the .10 level. Clicks within and after stressed words were identified more accurately than within and after unstressed words. This, too, is a significant difference, with the significance increasing from the .05 level

for position within the test word to the .01 level for position after the test word. The word Bill, however, shows the opposite result. In the case of Bill, the relationships between the scores are reversed, although only the difference between the scores for position within stressed and unstressed versions of Bill reaches significance (at the .01 level).

The various kinds of subjective shifts are shown in Table 5.

TABLE 5
CLICK PLACEMENT AND CLICK LOCATION IN STRESSED AND UNSTRESSED WORDS
(per cent)

Objective click placement	Subjective click location				
	Within preceding word	Before test word	Within test word	After test word	Within following word
Before did, stressed	8.0	16.0	46.0	8.0	2.0
Within did, stressed		10.0	68.0	16.0	6.0
After did, stressed		2.0	20.0	62.0	14.0
Before did, unstressed	11.3	26.7	50.7	2.0	0.7
Within did, unstressed	2.0	14.0	56.0	16.7	6.0
After did, unstressed	0.7	6.7	20.0	32.7	16.7
Before Bill, stressed	10.0	56.0	26.0	4.0	
Within Bill, stressed			38.0	48.0	6.0
After Bill, stressed		2.0	8.0	24.0	16.0
Before Bill, unstressed	9.3	50.0	26.0	8.7	0.7
Within Bill, unstressed	0.7	2.7	64.7	25.3	4.0
After Bill, unstressed			7.3	35.3	45.3
Before I, stressed	2.0	16.0	64.0	8.0	8.0
Within I, stressed		2.0	40.0	44.0	12.0
After I, stressed		2.0	20.0	66.0	12.0
Before I, unstressed	10.0	34.7	24.7	6.7	6.7
Within I, unstressed	2.0	17.3	24.0	22.0	28.0
After I, unstressed	1.3	9.3	3.3	33.3	50.7
Before thank, stressed	2.0	28.0	64.0	2.0	
Within thank, stressed	2.0	4.0	78.0		
After thank, stressed		2.0	36.0	22.0	20.0
Before thank, unstressed	9.3	46.0	33.3		
Within thank, unstressed	6.6	29.3	43.3	4.0	
After thank, unstressed	3.3	18.0	53.3	14.7	4.7

Study of this table explains why clicks preceding stressed words received low correct scores: there is an overwhelming tendency for such clicks to be subjectively located within the stressed word. To put it differently, stress attracts the click from the preceding boundary into the stressed word. For did, correct identification of a click before the test word was 16%, compared to subjective shifts in 46% of the cases; for I, the 16% correct location of the click occurring at the boundary contrasts with a 64% shift into the stressed word, and for thank, 28% correct contrasts with a 64% shift. The subjective shift in the case of I is particularly noteworthy, since it involves a shift away from the major syntactic boundary, which supposedly attracts clicks and certainly should resist their being attracted away. Table 6 shows the level of significance of differences in scores due to some of the shifts.

TABLE 6
DEGREE OF SIGNIFICANCE OF SUBJECTIVE SHIFTS

Objective Click Placement	Subjective shift (by one-half step) to	
	Within test word	Within following word
Before did, stressed	< .01	
After did, stressed		< .001
Before did, unstressed	< .01	
After did, unstressed		> .20
Before Bill, stressed	< .001	
After Bill, stressed		> .20
Before Bill, unstressed	< .01	
After Bill, unstressed		> .20
Before I, stressed	< .001	
After I, stressed		< .001
Before I, unstressed	> .20	
After I, unstressed		< .05
Before thank, stressed	< .001	
After thank, stressed		> .20
Before thank, unstressed	< .20	
After thank, unstressed		> .20

Table 6 requires some interpretation. It is obvious that the shifts from before a stressed word into the stressed word are highly significant. In some instances, shifts from after the

test word to the following word are also significant; but failure to shift is equally important. This is not shown directly on this table, but can be realized by comparing Table 6 with Table 5. For example, the probability that a click objectively placed after stressed did would be attracted into the following word is exceedingly small; the reason is the high accuracy of click location in that position in general, and the fact that no stressed word ever followed did. It is the stressed words that attract preceding clicks; there was no comparable systematic tendency for clicks to be subjectively shifted from a preceding boundary to the middle of an unstressed word.

As regards the word Bill, the degree of significance shows the failure to shift in both cases in which the click was placed before the word.

Clicks objectively placed within a stressed word receive high correct scores and show little tendency to shift away. This tendency is greater in unstressed words. The direction of these shifts is not systematic in any way.

Clicks placed after stressed words are highly identifiable. If they migrate, it is toward the following word. The tendency to shift into the following word is much more pronounced in the case of clicks placed after unstressed words. After Bill and I, in particular, the click was subjectively shifted to the following word more frequently than it was correctly located. Interestingly, this is the only instance in which unstressed Bill shares the behavior of other unstressed words; in all other respects, it seems as if stress and lack of stress were reversed in the case of Bill. The reason why clicks are not shifted to the following word after unstressed did and thank is most probably the lack of stress on the words immediately following the click.

Except for the matter just described, no particular regularities seemed to be associated with the position of the word relative to the beginning or end of the sentence. The behavior of clicks associated with Bill remains a problem calling for further study.

The results of the experiment demonstrate that stress does indeed have an effect on the subjective location of clicks. Without trying to read too much into the outcome of the limited experiment, I feel justified in saying that click localization is more sensitive to surface phenomena than as been previously assumed. The underlying structure of the sentence remained the same during the experiment; if the listeners somehow proceed directly to the analysis of underlying structures, clicks should have been treated similarly in the same words, regardless of their stressed or unstressed realization. Since there were significant differences, one may conclude that click localization is not exclusively dependent on the underlying syntactic structure of the sentence.

4. Summary and Conclusion.

In this paper, I have attempted to establish the units of perception and the levels at which perception operates. Evidence has been adduced for two basic steps in perception: primary processing and linguistic processing. Primary processing consists of auditory processing and phonetic processing, which constitutes listening in a speech mode. There are several levels within the linguistic level, of which the phonological and syntactic level are considered better documented than a possible morphological level. Linguistic processing presupposes primary processing. Auditory processing must logically precede other levels of processing; phonetic processing is considered as presupposed by the other levels, but the possibility is admitted that phonetic and linguistic processing may proceed concurrently. The units at the various levels may differ in size, and there is extensive interaction between them, as there is, for example, between the phonetic and phonological levels on the one hand and the syntactic level on the other hand. Processing at the syntactic level presupposes analysis at the phonetic level, which seems to be largely suprasegmental. Parallel processing is accepted as part of the model, and a strict separation of levels is considered unwarranted.

Footnote

*I am grateful to the College of Humanities of The Ohio State University for releasing me from teaching duties during the autumn quarter of 1971, while this paper was being written. I wish also to express my appreciation to Dr. P. B. Denes and Dr. J. P. Olive of the Bell Telephone Laboratories for their help with the experimental part of this paper, to Dr. A. W. F. Huggins (of M.I.T.) and Dr. T. Smith (of the University of California, San Diego) for their challenges and suggestions, and to my research assistants Linda R. Shockey and Richard P. Gregorski for their help in administering the listening test. This paper was presented at the April 1972 Vancouver symposium on "Speech Production--Speech Perception: Their Relationship to Cortical Functioning".

References

Abbs, J. H., and H. M. Sussman (1971) "Neurophysiological feature detectors and speech perception: a discussion of theoretical implications." JSHR 14.23-36.

Abrams, Kenneth, and Thomas G. Bever (1969) "Syntactic structure modifies attention during speech perception and recognition." Quarterly Journal of Experimental Psychology 21.280-290.

Bever, T., R. Kirk, and J. Lackner (1969) "An autonomic reflection of syntactic structure." Neuropsychologia 7.23-28.

Bever, T. G., J. R. Lackner, and R. Kirk (1969) "The underlying structures of sentences are the primary units of immediate speech processing." Perception and Psychophysics 5.225-234.

Bever, T. G., J. Lackner, and W. Stolz (1969) "Transitional probability is not a general mechanism for the segmentation of speech." Journal of Experimental Psychology 79.387-394.

Bond, Z. S. (1971) "Units in speech perception." Working Papers in Linguistics No. 9, viii-112. Computer and Information Science Research Center Technical Report Series, OSU-CISRC-TR-71-8. The Ohio State University, Columbus, Ohio.

Bondarko, L. V., N. G. Zagorujko, V. A. Kozhevnikov, A. P. Molchanov, and L. A. Chistovich (1968) "A model of speech perception by humans." Academy of Sciences of the U.S.S.R., Siberian Section: Nauka, Novosibirsk. Translated by I. Lehiste, Working Papers in Linguistics No. 6, Ohio State University, Columbus (1970) 88-132.

Chapin, Paul G., Timothy S. Smith, and Adele A. Abrahamson (1972, in press) "Two factors in perceptual segmentation of speech." Journal of Verbal Learning and Verbal Behavior.

Chistovich, L., G. Fant, A. de Serpa-Leitão, and P. Tjernlund (1966a) "Mimicking of synthetic vowels." Speech Transmission Laboratory Quarterly Progress and Status Report No. 2.1-18.

Chistovich, L., G. Fant, and A. de Serpa-Leitão (1966b) "Mimicking and perception of synthetic vowels, Part II." Speech Transmission Laboratory Quarterly Progress and Status Report 3.1-3.

Chistovich, L. A., and V. A. Kozhevnikov (1969) "Perception of Speech." in Voprosy teorii i metodov issledovanija vosprijatija recevyx signalov, Leningrad; Translated as L. A. Chistovich et al. "Theory and methods of research on perception of speech signals." JPRS 50423 (1970).

Day, Ruth S. (1970a) "Temporal order judgments in speech: are individuals language-bound or stimulus-bound?" (Paper presented at the 9th Annual Meeting of the Psychonomic Society, St. Louis, November, 1969). Haskins Laboratories Status Report SR-21/22, 71-87.

Day, Ruth S. (1970b) "Temporal order perception of a reversible phoneme cluster." Paper presented at the 79th meeting of the Acoustical Society of America, Atlantic City, 21-24 April.

Day, Ruth S., and James E. Cutting (1970) "Perceptual competition between speech and nonspeech." Paper presented at the 80th meeting of the Acoustical Society of America, Houston, 3-6 November.

Denes, Peter B. (1963) "On the statistics of spoken English." JASA 35.892-904.

Denes, Peter B. (1970) "On-line computers for speech research." Transactions of the IEEE on Audio- and Electroacoustics, December, Vol. AU-18, No. 4, 418-425.

Fodor, J. A., and T. G. Bever (1965) "The psychological reality of linguistic segments." Journal of Verbal Learning and Verbal Behavior 4.414-420. Also in: L. A. Jakobovits and M. S. Miron (eds.), Readings in the Psychology of Language. Englewood Cliffs, N.J.: Prentice-Hall, Inc. (1964) 325-332.

Fodor, J. A., and M. F. Garrett (1971) "A consolidation effect in sentence perception." M.I.T. Research Laboratory of Electronics Quarterly Progress Report No. 100, January 15, 182-185.

Fry, D. B. (1970) "Reaction time experiments in the study of speech processing." Nouvelles Perspectives en phonétique, Institut de Phonétique, Université Libre de Bruxelles: Conferences et Travaux, Vol. 1, 15-35.

Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman (1962) "The identification and discrimination of synthetic vowels." Language and Speech 5.171-189.

Fujisaki, H., and T. Kawashima (1968) "The influence of various factors on the identification and discrimination of synthetic speech sounds." Reports of the 6th International Congress on Acoustics, Tokyo, No. 2.B-95-98.

Fujisaki, H., and T. Kawashima (1969) "On the modes and mechanisms of perception of speech sounds." Paper presented at the 78th meeting of the Acoustical Society of America, San Diego, November 4.

Gamlin, Peter J. (1971) "Sentence processing as a function of syntax, short term memory capacity, the meaningfulness of the stimulus and age." Language and Speech 14.115-134.

Garrett, M., T. Bever, and J. Fodor (1965) "The active use of grammar in speech perception." Perception and Psychophysics 1.30-32.

Harris, Z. (1944) "Simultaneous components in phonology." Language 20.181-205.

Holmes, V., and K. Forster (1970) "Detection of extraneous signals during sentence recognition." Perception and Psychophysics 7.5.297-301.

Johnson, Neal F. (1970) "The role of chunking and organization in the process of recall." Psychology of Learning and Motivation, Vol. 4, Academic Press, Inc.: New York, 171-247.

Ladefoged, P., and D. E. Broadbent (1960) "Perception of sequence in auditory events." Quarterly Journal of Experimental Psychology 12.162-170.

Lane, H. (1965) "The motor theory of speech perception: a critical review." Psychological Review 2.275-309.

Lehiste, Ilse (1967) "Suprasegmental features, segmental features, and long components." Actes du Xe congrès international des linguistes, Bucarest, 1967: Editions de l'academic de la Republique socialiste de Roumanie, Bucarest, Vol. IV.1-7 (1970).

Lehiste, Ilse (1970a) Suprasegmentals. M.I.T. Press:Cambridge.

Lehiste, Ilse (1970b) "Experiments with synthetic speech concerning quantity in Estonian." Proceedings of the 3rd International Congress of Fenno-Ugricists, Tallinn (in press).

Lehiste, Ilse, and L. Shockey (1971) "The perception of coarticulation." Two papers presented at the 82nd meeting of the Acoustical Society of America, Denver, October 20.

Liberman, Alvin M. (1957) "Some results of research on speech perception." JASA 29.117-123.

Liberman, Alvin M. (1970) "The grammars of speech and language." Cognitive Psychology 1.301-323.

Liberman, A. M., F. S. Cooper, K. S. Harris, and P. F. MacNeilage (1962) "A motor theory of speech perception." Proceedings of Speech Communication Seminar, Stockholm, Session D-3, 1-10.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967) "Perception of the speech code." Psychological Review 74.431-461.

Liberman, A. M., K. S. Harris, N. Hoffman, and B. Griffith (1957) "The discrimination of speech sounds within and across phoneme boundaries." Journal of Experimental Psychology 54.358-368.

Liberman, A. M., K. S. Harris, J. Kinney, and H. Lane (1961) "The discrimination of relative onset time of the components of certain speech and nonspeech patterns." Journal of Experimental Psychology 61.379-388.

Lisker, L., and A. S. Abramson (1971) "Distinctive features and laryngeal control." Language 47.767-785.

Miller, G. A. (1956) "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review 63.81-97.

Miller, G. A. and P. E. Nicely (1955) "An analysis of perceptual confusions among some English consonants." JASA 27.338-352.

Neisser, Ulric (1967) Cognitive Psychology. New York: Appleton-Century-Crofts.

Öhman, S. E. G. (1966) "Coarticulation in VCV utterances." JASA 39.151-168.

Olive, J. P. (1971) "Automatic formant tracking by a Newton-Raphson technique." JASA 50.661-670.

Pisoni, David B. (1971) "Very brief short-term memory in speech perception." Paper presented at the 82nd meeting of the Acoustical Society of America, Denver, October 19.

Rabiner, L. R., et al. (1971) "Digital formant synthesis." Paper 23C8, Proceedings of the 7th International Congress on Acoustics, Budapest, Vol. 3.157-158.

Savin, H. B., and T. G. Bever (1970) "The nonperceptual reality of the phoneme." Journal of Verbal Learning and Verbal Behavior 9.295-302.

Sharf, Donald J. (1971) "Perceptual parameters of consonant sounds." Language and Speech 14.169-177.

Spiegel, Murray R. (1961) Theory and Problems of Statistics. McGraw-Hill: New York, 171.

Stevens, K. N., A. M. Liberman, S. E. G. Öhman, and M. Studdert-Kennedy (1969) "Cross-language study of vowel perception." Language and Speech 12.1-23.

Studdert-Kennedy, Michael, and Donald Shankweiler (1970) "Hemispheric specialization for speech perception." JASA 48.579-594.

Studdert-Kennedy, Michael, A. M. Liberman, K. S. Harris, and F. D. Cooper (1970) "Motor theory of speech perception: A reply to Lane's critical review." Psychological Review 77.234-249.

Wickelgren, Wayne A. (1969a) "Context-sensitive coding, associative memory, and serial order in (speech) behavior." Psychological Review 76.1.1-15.

Wickelgren, Wayne A. (1969b) "Context-sensitive coding in speech recognition, articulation and development." In K. N. Leibovic, ed., Information Processing in the Nervous System. Springer: New York-Heidelberg-Berlin, 85-95.